IBM Developer
SKILLS NETWORK

Winning Space Race with Data Science

João Gilberto Perna
22/05/2024

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion

Executive Summary

- Summary of methodologies
  - Data Collection
  - Webscraping
  - Data Wrangling
  - EDA with Data Visualization and SQL
  - Mapping with Folium
  - Interactive Dashboard with Plotly
  - Predictive model testing and Analysis
- Summary of all results
  - Data Analysis results
  - Predictive model evaluation and results

Introduction

- SpaceX has arguably become the most important player in the commercial space industry, using cost-cutting measures to sell launches of its Falcon 9 rocket on its website for $62 million, when its competitors are charging upwards of $165 million per launch. Among the most important cost-cutting measures is the ability to reuse the rocket's first stage. As a result, being able to predict the first stage's successful landing is an important step in being able to predict overall launch costs. Purpose: In our project, we will build models to predict reusability of the first stage, using publicly available data and the use of machine learning techniques.

Key questions to address:

- How do factors such as payload mass, launch site, flight frequency, and orbital paths influence the success of the first stage's landing ?
- Has there been an upward trend in the rate of successful landings over time ?
- Which algorithm is most effective for binary classification in this scenario ?

Section 1

Methodology

Methodology

Executive Summary

- Data collection obtained using :
  - SpaceX REST API https://api.spacexdata.com/v4/rockets
  - Web Scrapping using BeautifulSoup https://en.wikipedia.org/wiki/List_of_Falcon/_9/_and_Falcon_Heavy_launches
- Perform data wrangling
  - Utilization of Python libraries pandas and numpy for data manipulation.
  - Application of One Hot Encoding for the development of classification models.
- Perform exploratory data analysis (EDA) using visualization and SQL

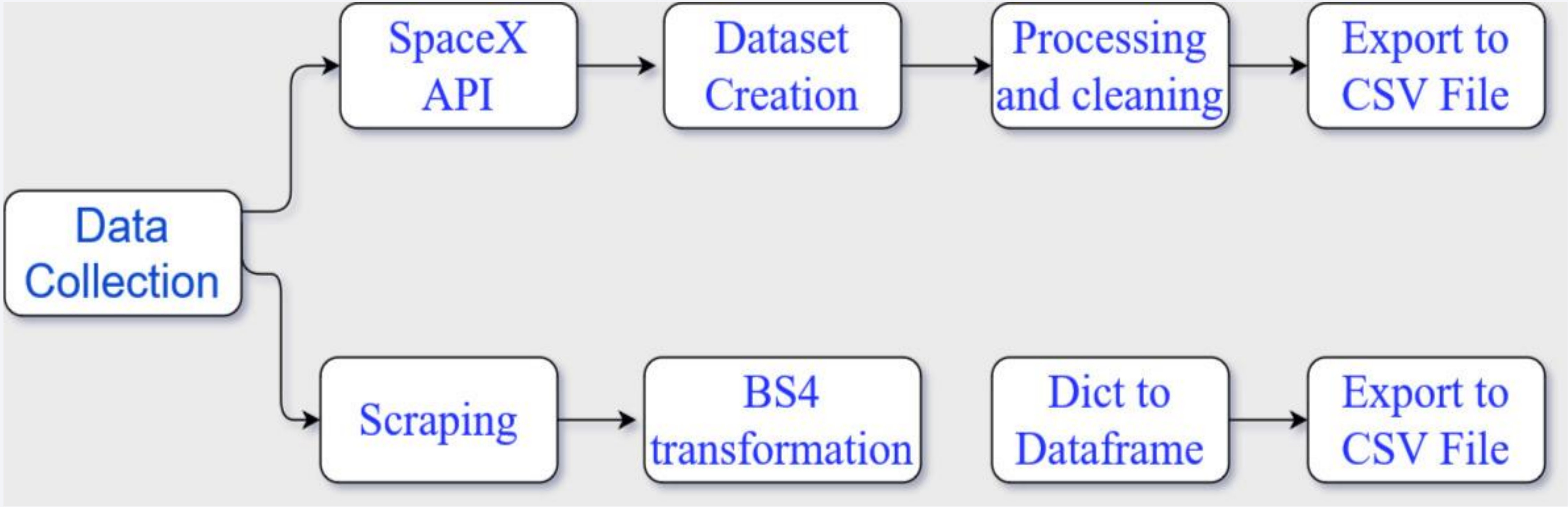
Methodology

Executive Summary

- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Experiment usability and compatibility of SVM, Tree maps, KNN, Logistic regression
  - Parameter optimization was assessed using sklearn

Data Collection

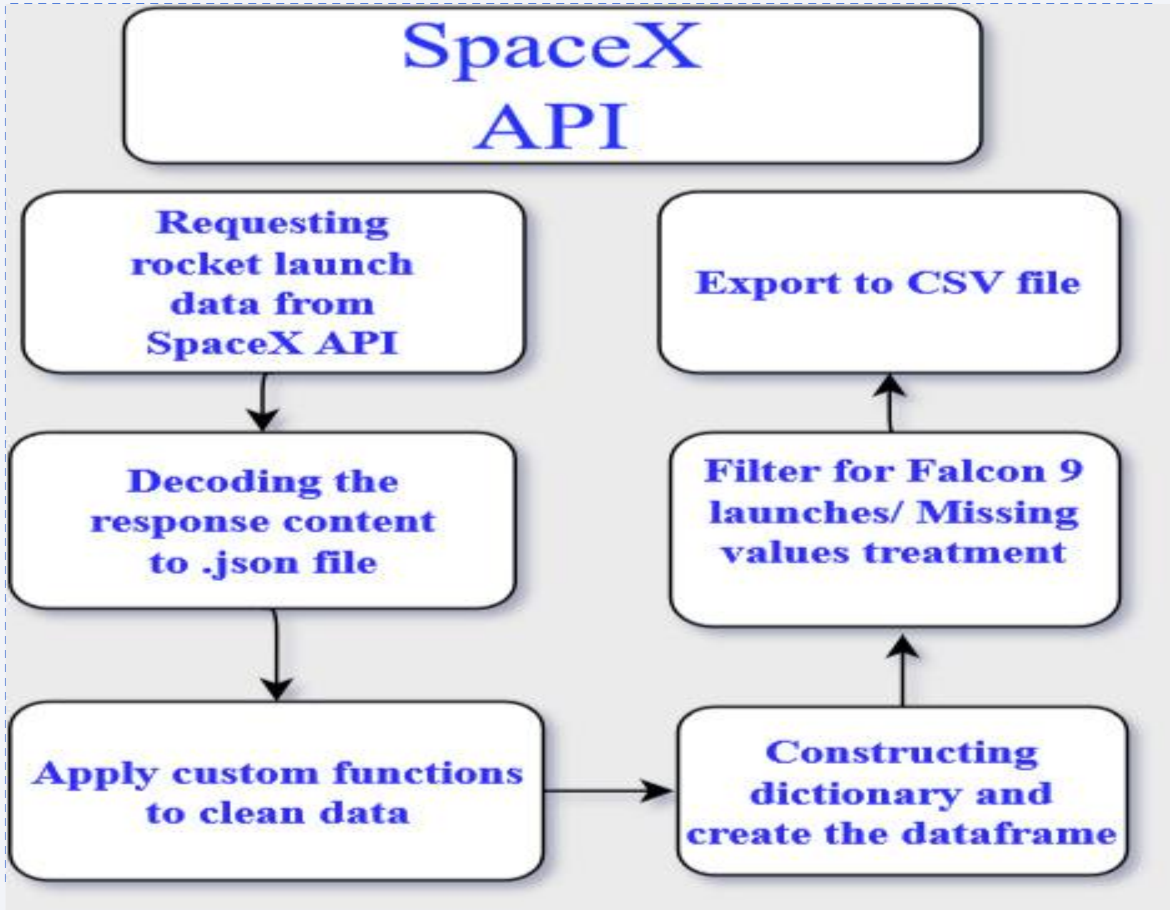
- The data was collected:
    1. Using SpaceX API
    2. Using BeautifulSoup library to scrap data from Wikipedia



Data Collection – SpaceX API

- Extract from API
- Convert data
- Clean an filter
- Store data in a flat file

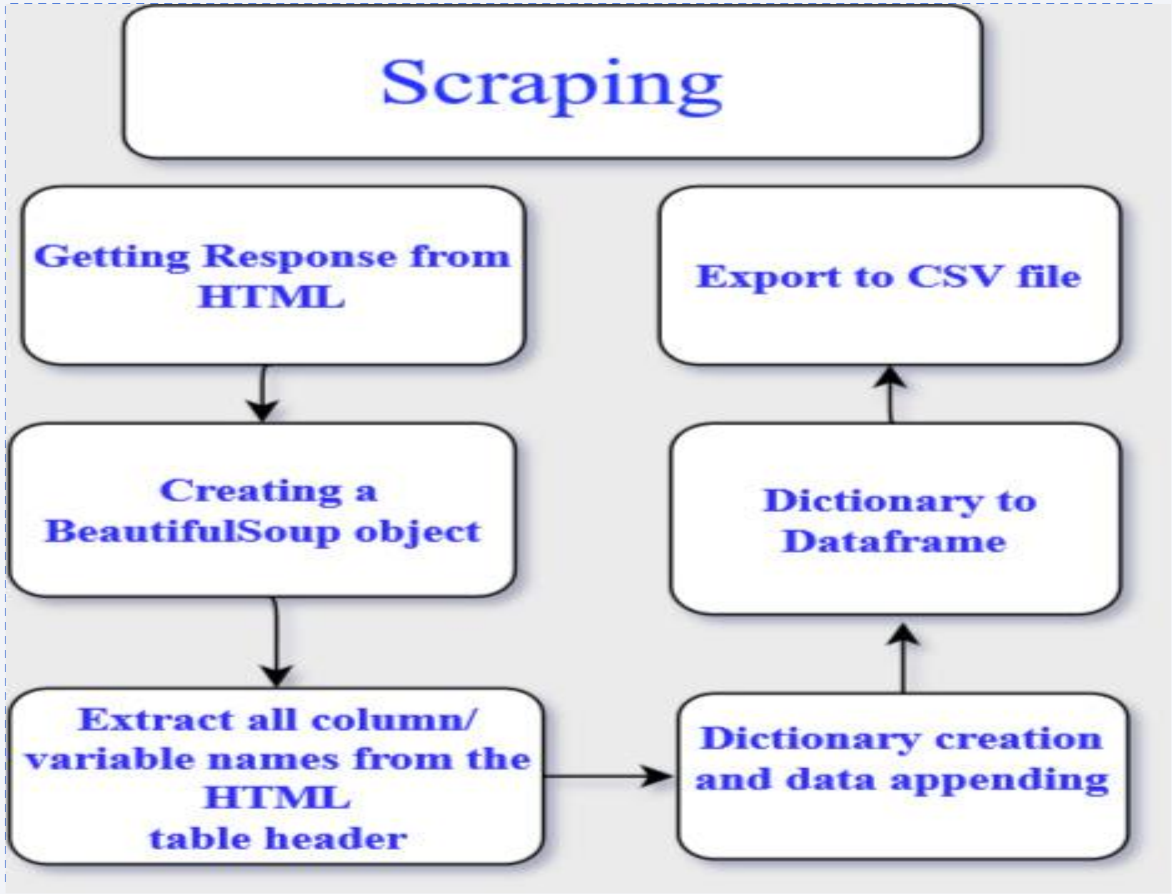
- Source Code:
  - GitHub



Data Collection - Scraping

- Extract from Wikipedia
- Parse table using beautifulSoup4 library
- Convert to structured data
- Store data in a flat file

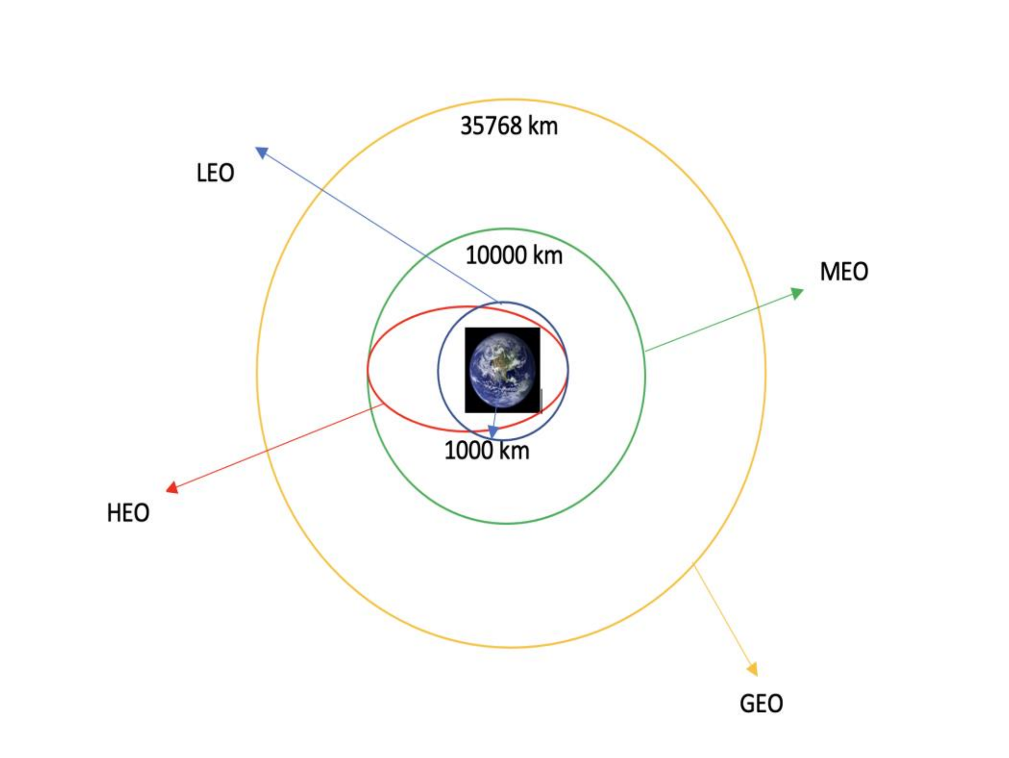
- Source Code:
  - [GitHub]()



Data Wrangling

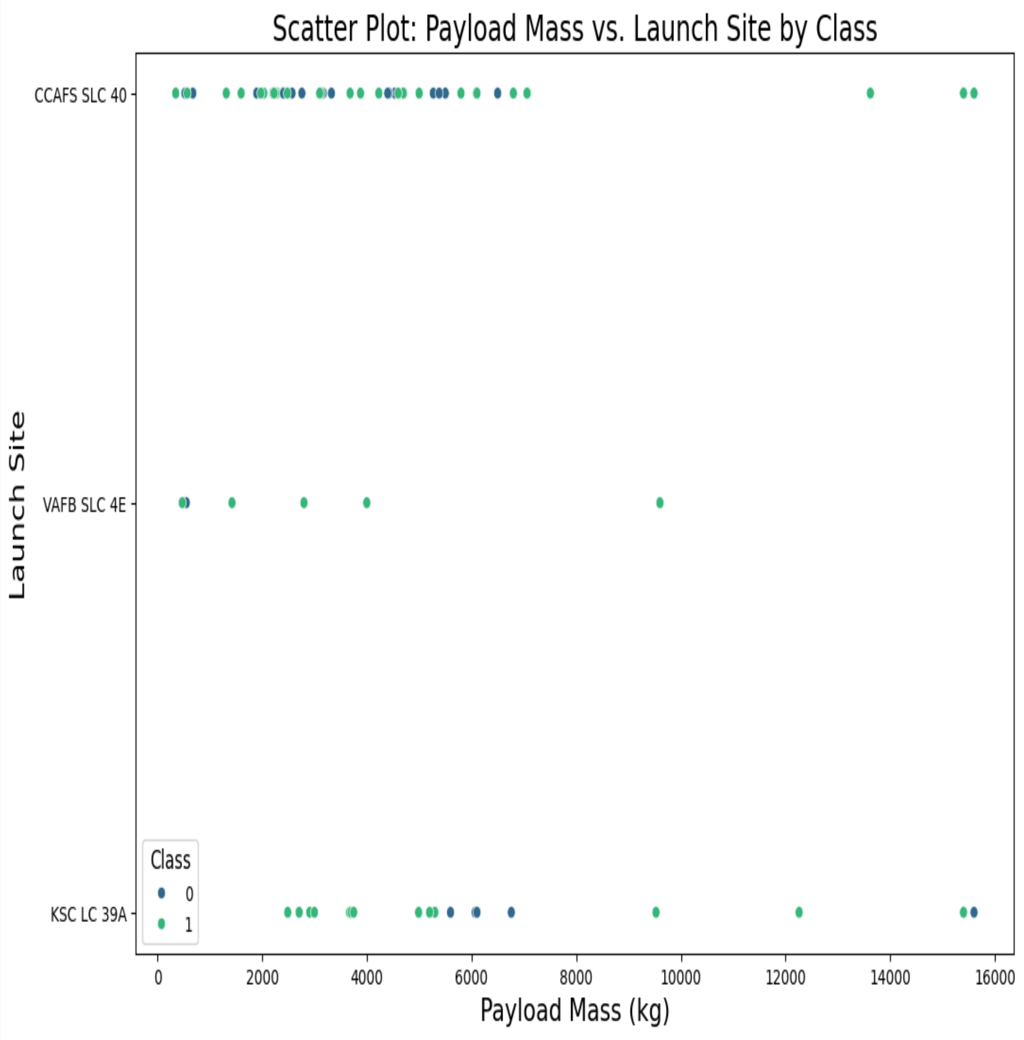
1. Exploratory Data Analysis (EDA) was performed on the Dataset
2. Calculate the number of launches on each site
3. Calculate the number of occurrences of outcome per orbit.
4. Create a landing outcome label from Outcome column
5. Export Data

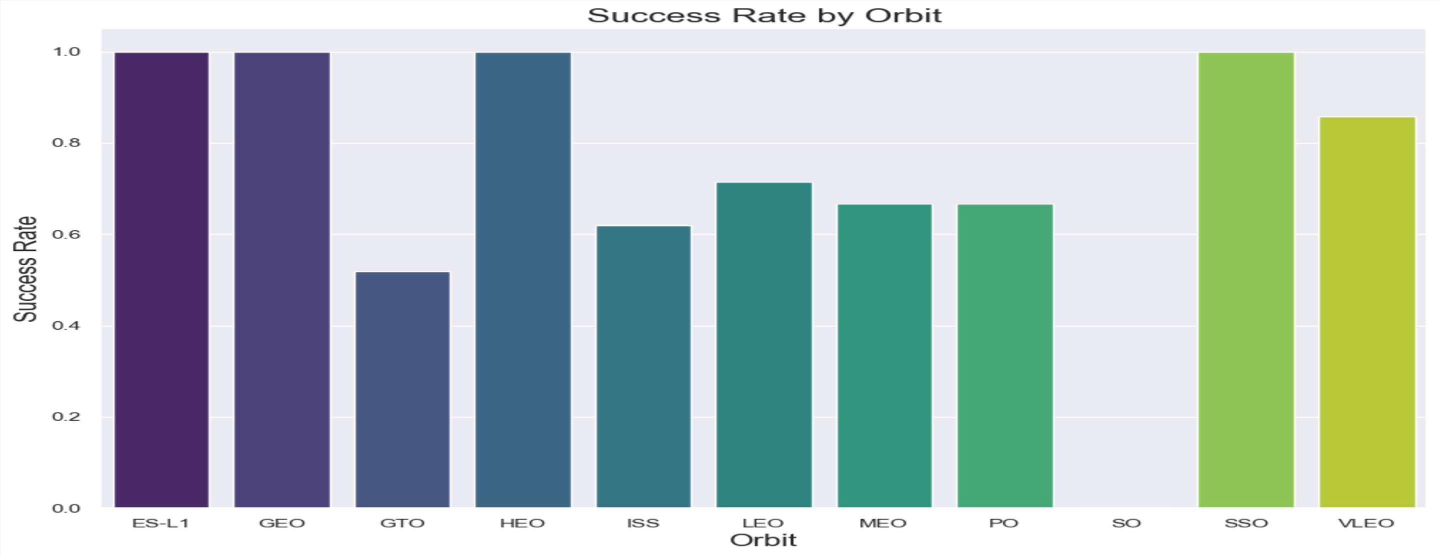
- Source Code:
    - [GitHub]()



Example of different orbit types

EDA with Data Visualization





[GitHub]

EDA with SQL

To gain deeper insights into the SpaceX dataset, the following SQL queries/operations were performed on an IBM DB2 cloud instance:

1. Display the names of the unique launch sites in the space mission
2. Display 5 records where launch sites begin with the string 'CCA'
3. Display the total payload mass carried by boosters launched by NASA (CRS)
4. Display average payload mass carried by booster version F9 v1.1
5. List the date when the first successful landing outcome in ground pad was acheived.
6. List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000
7. List the total number of successful and failure mission outcomes
8. List the names of the booster_versions which have carried the maximum payload mass. Use a subquery
9. List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015
10. Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.

GitHub

Build an Interactive Map with Folium

Map Objects	Code	Result
Map Marker	foluim.Marker()	Map object to make a mark on map.
Icon Marker	folium.Icon()	Create na icon on map
Circle Marker	folium.Circle()	Create a circle Where marker is placed
PolyLine	folium.Polyline()	Create a line in between points
Mouse Position	MousePosition()	Helps to find the coordinates easily of any points of interests while exploring the map
Marker Cluster Object	MarkerCluster()	Good way to simplify a map containing many markers having the same coordinate.

Folium interactive map helps analyze geospatial data to perform more interactive visual analytics and better understand factors such location and proximity of launch sites that impact launch success rate.

Key takeaways

- All launch site are close proximity to railways
- All launch site are close proximity to highways
- All launch site are close proximity to coastline
- All launch site keep a certain distance from cities

GitHub

Build a Dashboard with Plotly Dash

Built a Plotly Dash web application to perform interactive visual analytics on SpaceX launch data in real-time

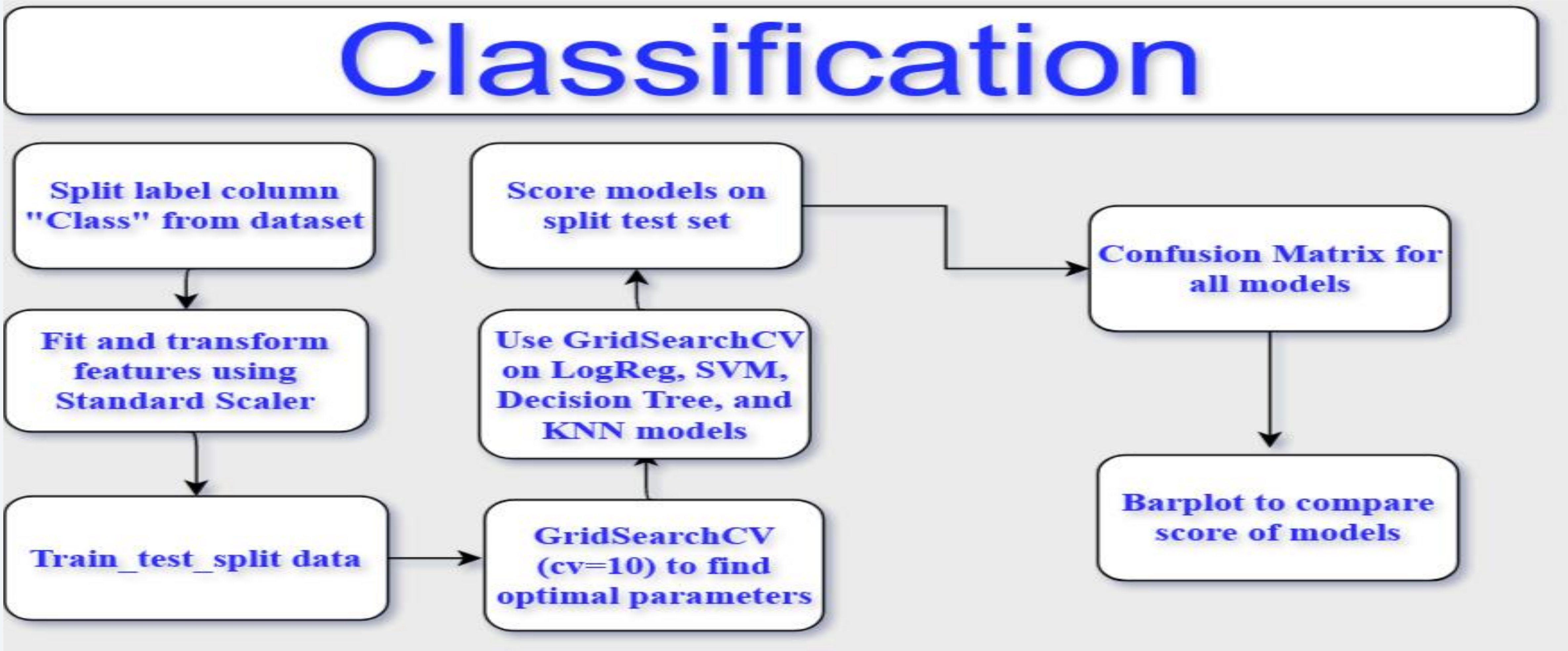
- Launch Site Dropdown Menu:
  - Implemented a dropdown menu for selecting launch sites.
- Success Launches Overview (All Sites/Specific Site):
  - Created a pie chart to visualize the total successful launches across all sites and the breakdown of success versus failure for a selected launch site.
- Payload Mass Range Slider:
  - Integrated a slider for selecting the payload mass range.
- Booster Versions Scatter Plot:
  - Developed a scatter plot illustrating the relationship between payload mass and launch success rate for various booster versions.

[GitHub]

Key takeaways

- KSC LC-39A has the largest successful launches
  - 10 in total with 76,9% success rate
- F9 Booster version with highest launch success rate is FT
- Payload range(s) with most highest launch success rate is 2000 – 5000 kg
- Payload range(s) with lowest launch success rate 0 – 2000 and 5500 – 7000

Predictive Analysis (Classification)



GitHub

Results

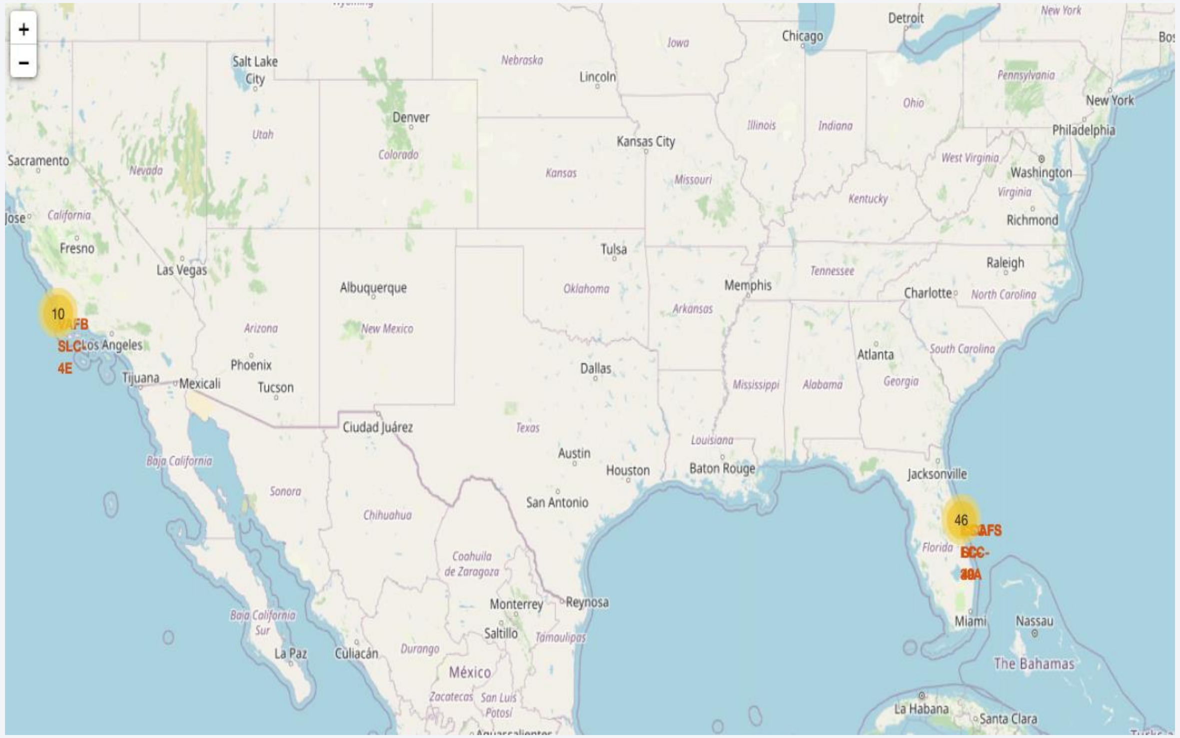
Exploratory data analysis results

- SpaceX utilizes four distinct launch locations.
- Initial missions were conducted jointly by SpaceX and NASA.
- The average payload of the F9 v1.1 booster is around 2,928 kilograms.
- In 2015, five years following the first launch, the initial successful booster landing occurred.
- Multiple Falcon 9 booster variants have landed successfully on drone ships, often with payloads exceeding the average weight.
- During 2015, two specific booster models, F9 v1.1 B1012 and F9 v1.1 B1015, did not achieve successful landings on drone ships.
- The rate of successful landings has increased over time.

Results

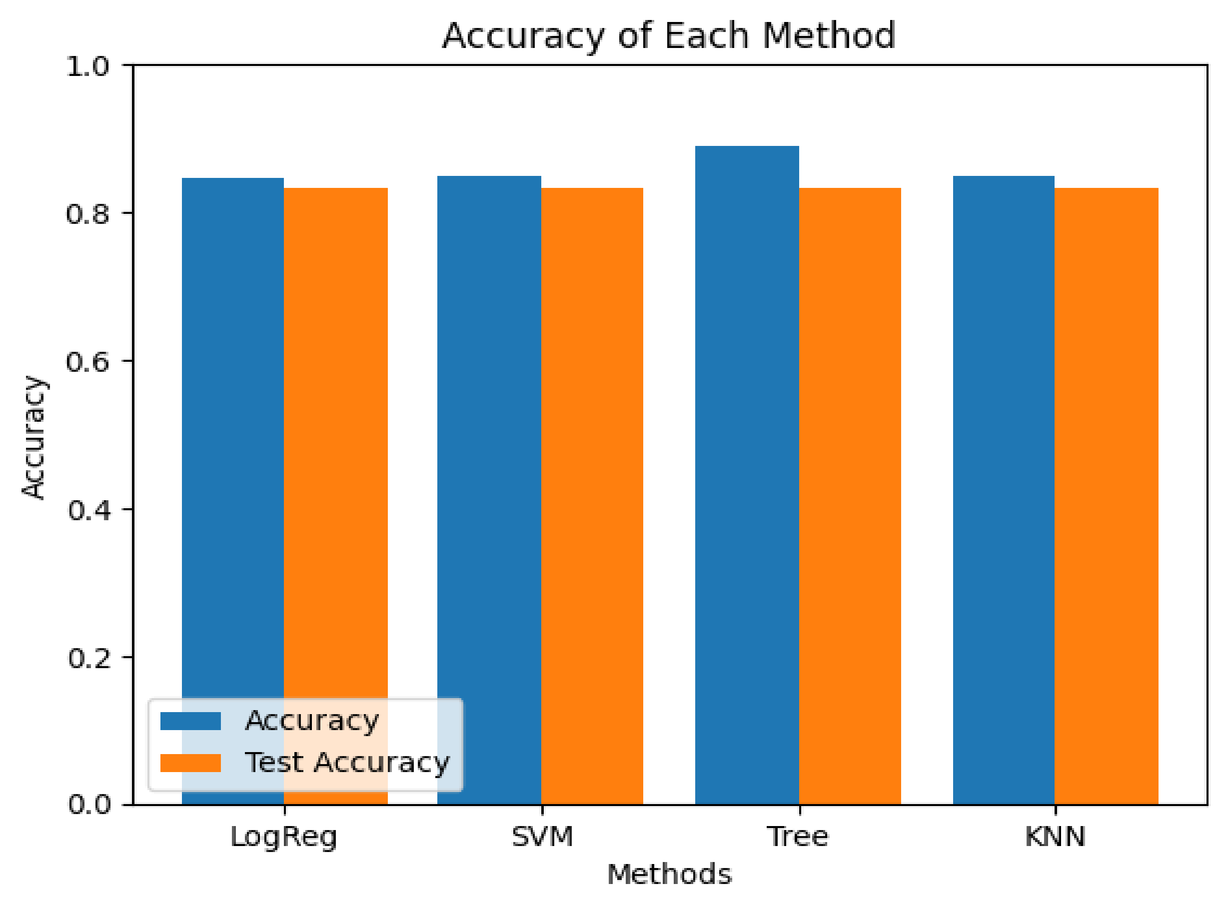
Interactive analytics demo in screenshots

- Interactive analytics reveal that launch sites are strategically located in secure areas near the sea with robust logistical support.
- The majority of launches take place from launch sites on the east coast.



Results

Predictive analysis results

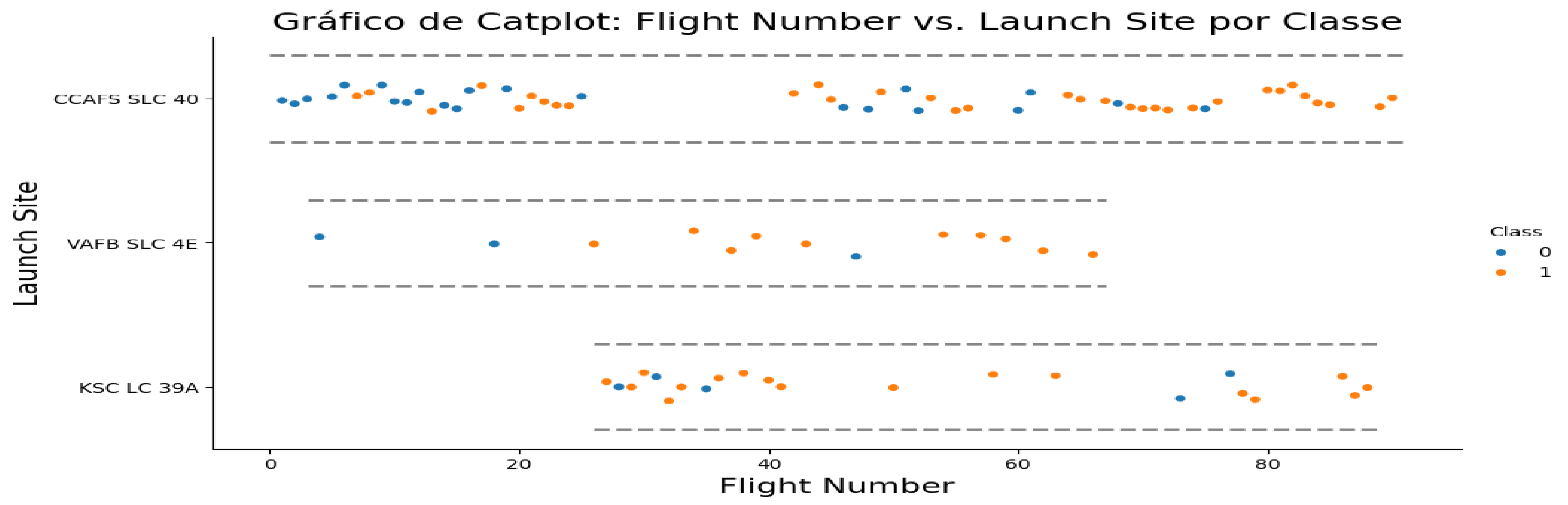


Predictive analysis indicates that the Decision Tree Classifier is the best model for predicting successful landings, with an accuracy exceeding 88%, compared to less than 85% accuracy from the other three models

Section 2

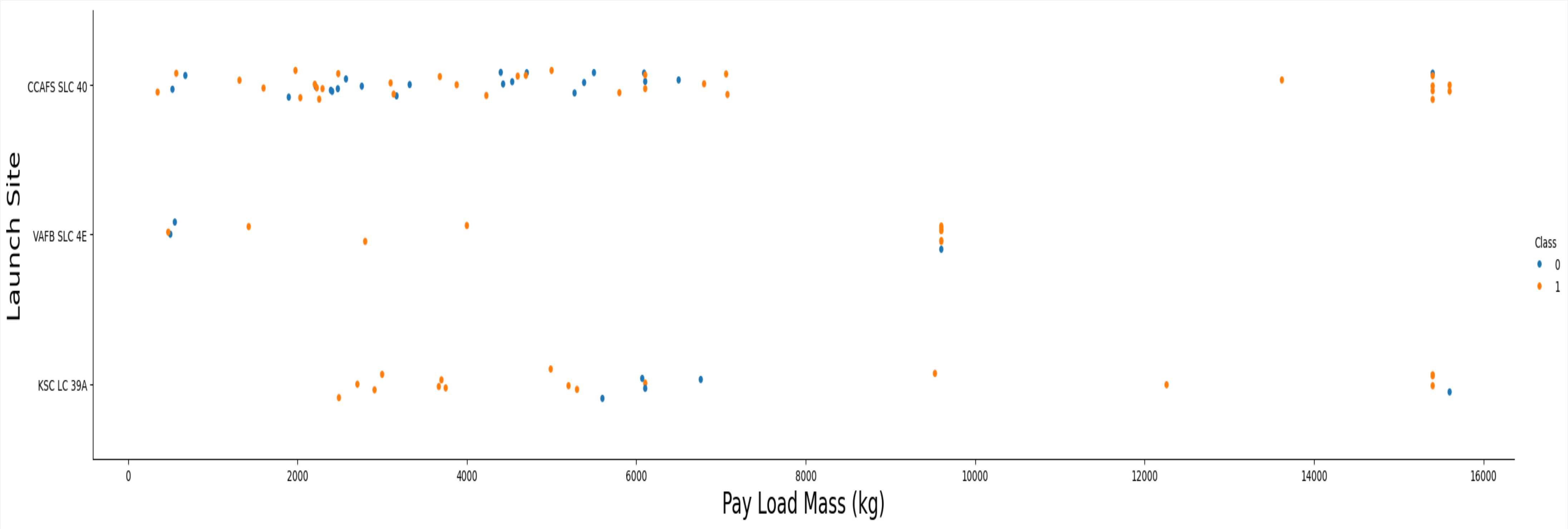
Insights drawn from EDA

Flight Number vs. Launch Site



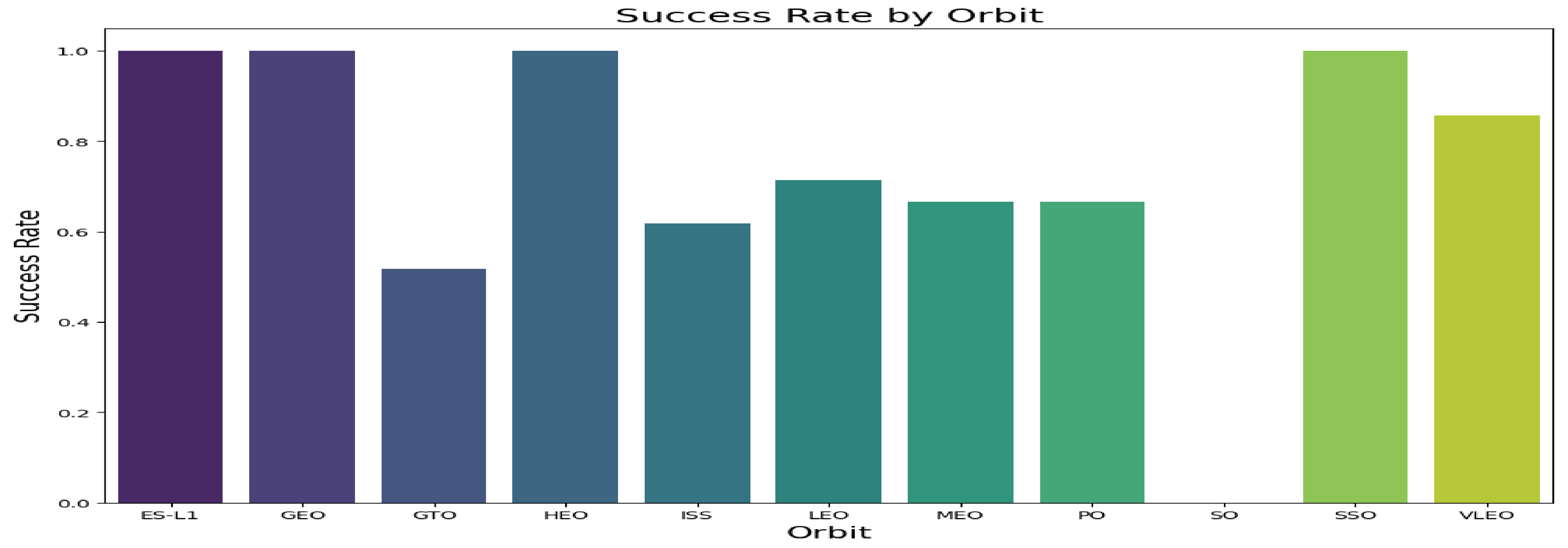
- According to the data, it's evident that CCAF5 SLC 40 has the highest number of successful launches currently.
- There's an increasing success rate over time.
- It's notable that after 30 flights, the success rate substantially rises.

Payload vs. Launch Site



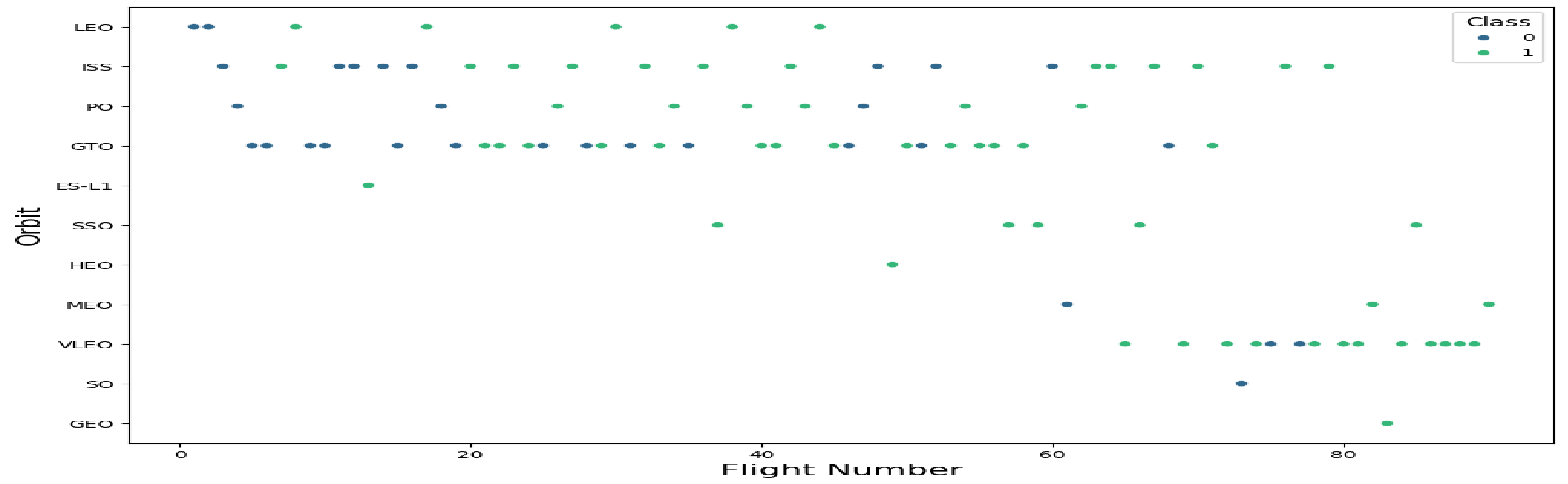
- The payload impact may indicate correlation in some cases, but not causality.
- Payloads exceeding 7,000 kg show an extremely high success rate.
- KSC LC 39A maintains a 100% success rate for payloads under 5500 kg

Success Rate vs. Orbit Type



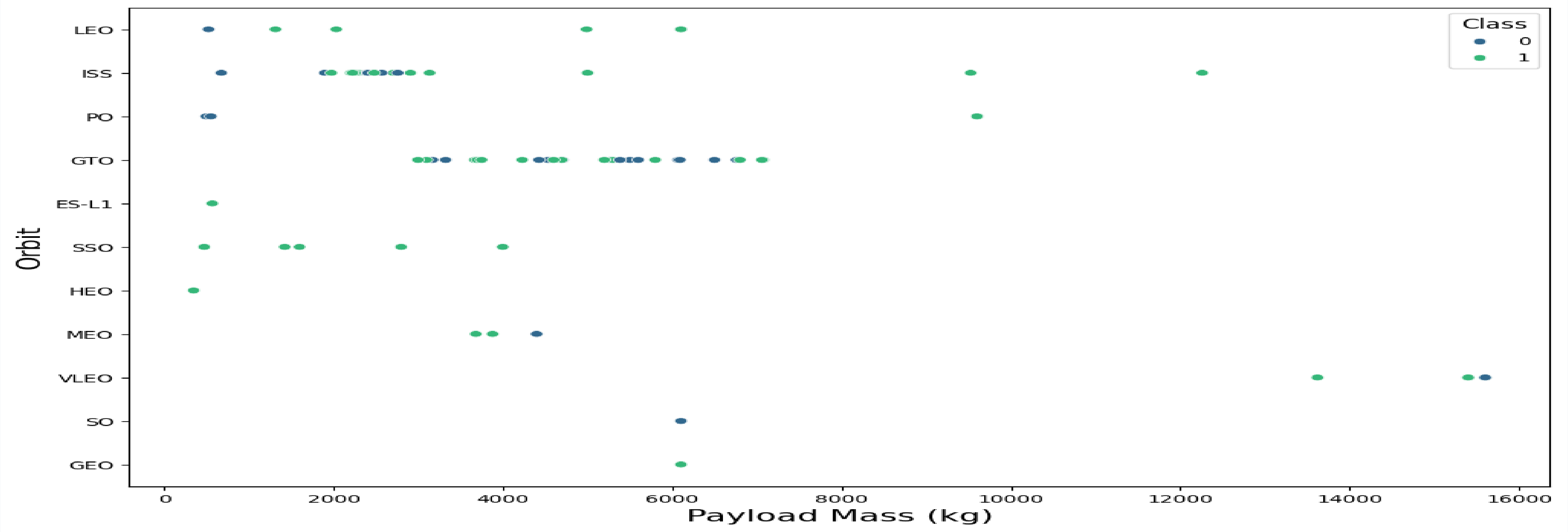
- Orbits ES-LI, GEO, HEO, and SSO have the highest success rates
- GTO orbit has the lowest success rate
- However, it's important to note that the number of observations for some orbit types is low because a 100% success rate is suspicious.

Flight Number vs. Orbit Type



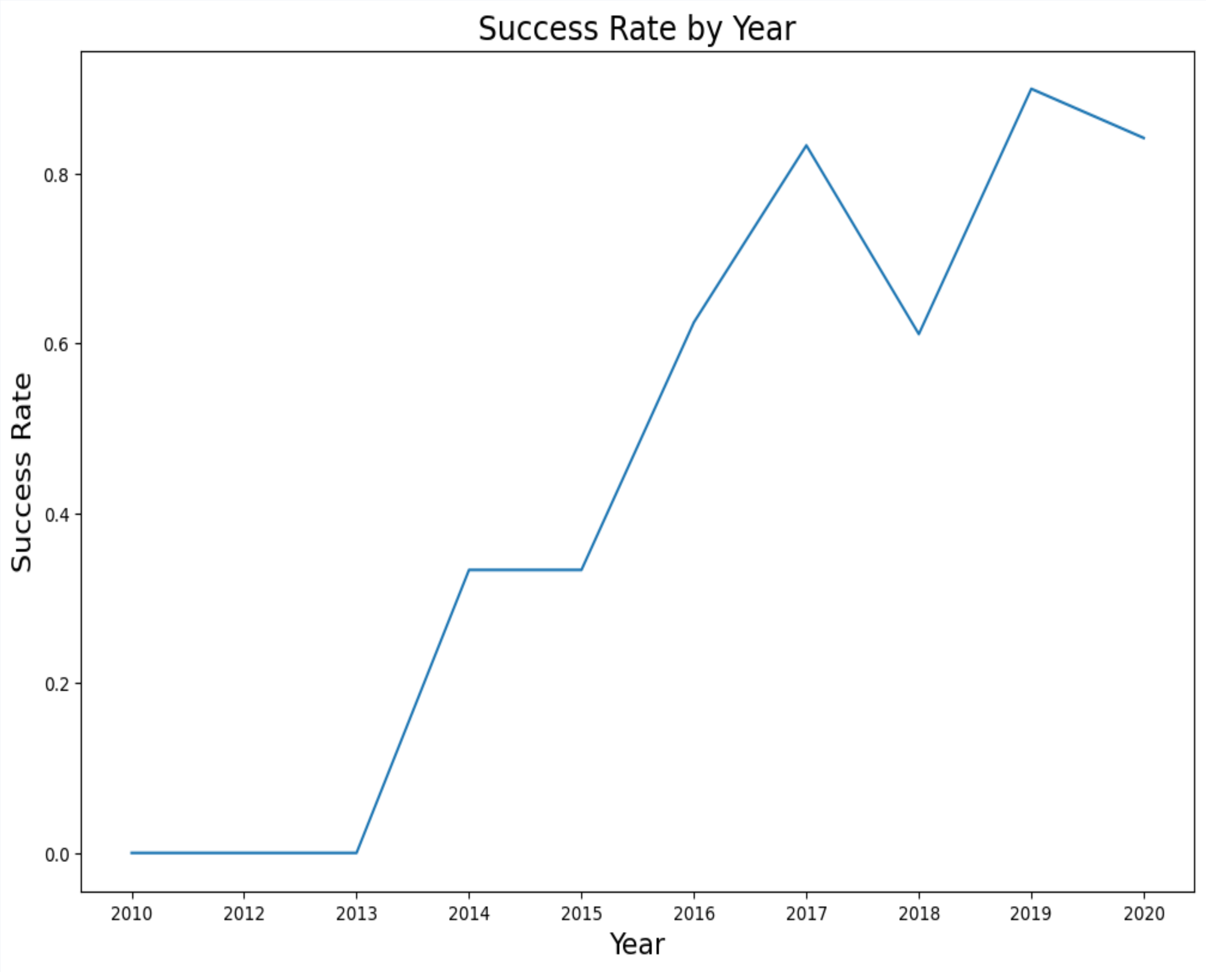
- For the majority of orbits (LEO, ISS, PO, SSO, MEO, VLEO), the likelihood of successful landings seems to rise with the number of flights
- It seems that there's no correlation between flight count and orbit type for GTO
- The VLEO orbit appears to present a new business opportunity, given its recent increase in frequency
- Some orbit types indeed have insufficient observations to be accurately measured

Payload vs. Orbit Type



- Heavy payloads have negative effect MEO, GTO, and VLEO orbits.“
- Positive effect for LEO and ISS orbits."

Launch Success Yearly Trend



- We can observe a period of failures between 2010 and 2013 (likely a testing and technology refinement phase).
- The success rate substantially increases from 2013 onwards

All Launch Site Names

SQL Query

```
%%sql
SELECT DISTINCT Launch_Site
FROM SPACEXTABLE
```

Description

- With ‘distinct’ we can return only unique values from the queries column (Launch_Site)
- There are 4 unique launch sites

Launch_Site
CCAFS LC-40
VAFB SLC-4E
KSC LC-39A
CCAFS SLC-40

Launch Site Names Begin with 'CCA'

Date	Time (UTC)	Booster_Version	Launch_Site	Payload	PAYLOAD_MASS__KG_	Orbit	Customer	Mission_Outcome	Landing_Outcome
2010-04-06	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-08-12	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-08-10	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-01-03	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

SQL Query

```
%%sql
SELECT *
FROM SPACEXTABLE
WHERE Launch_Site LIKE 'CCA%'
LIMIT 5;
```

Description

Displaying 5 records where launch sites begin with the string 'CCA'

Total Payload Mass

SQL Query

```
%%sql
SELECT SUM(PAYLOAD_MASS__KG_)
FROM SPACEXTABLE
WHERE Customer = "NASA (CRS)"
```

SUM(PAYLOAD_MASS__KG_)
45596

Description

Utilizing the SUM function, compute the aggregate in the PAYLOAD_MASS_KG column, applying a WHERE condition to select records where the Customer's name is 'NASA (CRS)'

Average Payload Mass by F9 v1.1

SQL Query

```
%%sql
SELECT AVG(PAYLOAD_MASS__KG_)
FROM SPACEXTABLE
WHERE Booster_Version = 'F9 v1.1';
```

AVG(PAYLOAD_MASS__KG_)
2928.4

Description

Utilizing the AVG function to determine the mean value in the PAYLOAD_MASS_KG column, using a WHERE clause to restrict the analysis to entries with the Booster_version 'F9 v1.1'

First Successful Ground Landing Date

SQL Query

```
%%sql
SELECT MIN(Date)
FROM SPACEXTABLE
WHERE Landing_Outcome LIKE '%Success%'
```

Description

Apply the MIN function to identify the earliest date in the Date column, using a WHERE clause to limit the analysis to records where the Landing_Outcome contain the word 'Success'

MIN(Date)
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

SQL Query

```
%%sql
SELECT Booster_Version
FROM SPACEXTABLE
WHERE Landing_Outcome =  'Success (drone ship)' AND
PAYLOAD_MASS__KG_ > 4000 AND PAYLOAD_MASS__KG_ < 6000;
```

Description

Select exclusively the Booster_Version, applying a WHERE clause to narrow the dataset to records where the Landing_Outcome equals 'Success (drone ship)'. Additionally, the AND clause sets further conditions, specifying Payload_MASS_KG must be greater than 4000 and less than 6000.

Booster_Version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

SQL Query

```
%%sql
SELECT mission_outcome,
count(mission_outcome) as total_number
from SPACEXTBL
group by mission_outcome;
```

Description

Select exclusively the mission_outcome and its occurrence count to tally the total outcomes of missions classified as success, failure, or unclear, ultimately grouping the data by mission_outcome

Mission_Outcome	total_number
Success	99
Failure (in flight)	1
Success (payload status unclear)	1

Boosters Carried Maximum Payload

Booster_Version
F9 B5 B1048.4
F9 B5 B1049.4
F9 B5 B1051.3
F9 B5 B1056.4
F9 B5 B1048.5
F9 B5 B1051.4

Booster_Version
F9 B5 B1049.5
F9 B5 B1060.2
F9 B5 B1058.3
F9 B5 B1051.6
F9 B5 B1060.3
F9 B5 B1049.7

Description

Using the DISTINCT function, unique values of Booster_Version can be obtained. Subsequently, a subquery is performed to select only those records where the PAYLOAD_MASS_KG is at its maximum

SQL Query

```
%%sql
SELECT DISTINCT(Booster_Version)
FROM SPACEXTBL
WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEXTBL);
```

2015 Launch Records

Date	Booster_Version	Launch_Site	Landing_Outcome
2015-01-10	F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
2015-04-14	F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

SQL Query

```
%%sql
SELECT
date,
BOOSTER_VERSION,
LAUNCH_SITE,
landing_outcome
FROM SPACEXTBL
WHERE DATE LIKE '2015-%' AND
Landing_Outcome = 'Failure (drone ship)';
```

Description

Select the date, Booster_Version, Launch_site, and landing_outcome, applying conditions to include only records from the year 2015 and where the outcome is a failure

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

Landing_Outcome	COUNT(Landing_Outcome)
No attempt	10
Success (drone ship)	5
Failure (drone ship)	5
Success (ground pad)	3
Controlled (ocean)	3
Uncontrolled (ocean)	2
Failure (parachute)	2
Precluded (drone ship)	1

SQL Query

```
%%sql
SELECT Landing_Outcome, COUNT(Landing_Outcome)
FROM SPACEXTBL
WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY  Landing_Outcome
ORDER BY COUNT(Landing_Outcome) DESC;
```

Description

Select Landing_Outcomes from the period between June 4, 2010, and March 20, 2017, ordering them in descending order

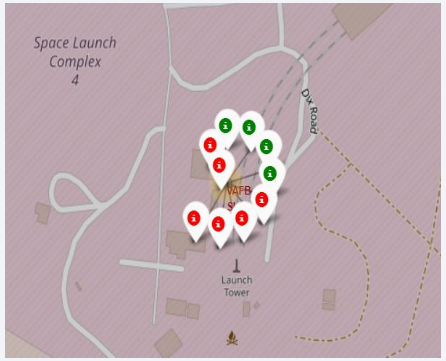
Section 3

Launch Sites Proximities Analysis

SpaceX Falcon9 - Launch Sites Map

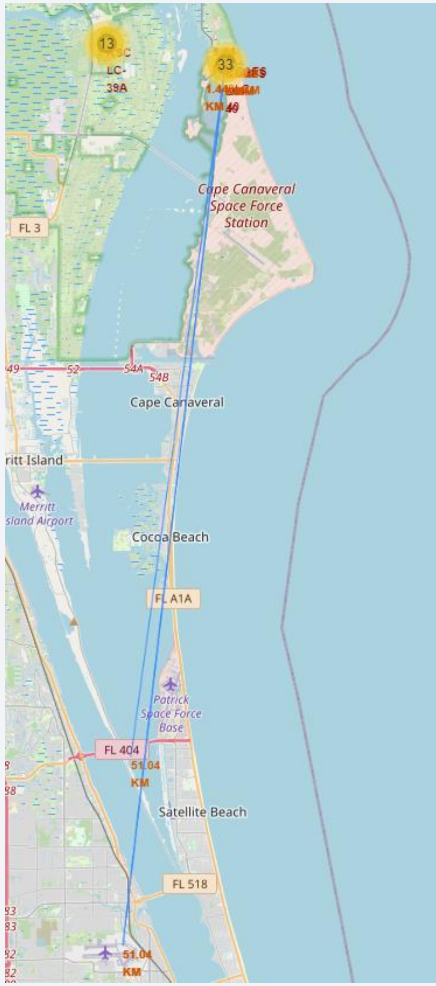
- All launch locations are situated near the coastline, particularly in the Florida and California regions, this ensur that when rockets are launched toward the ocean, it reduces the possibility of debris falling or exploding close to populated areas

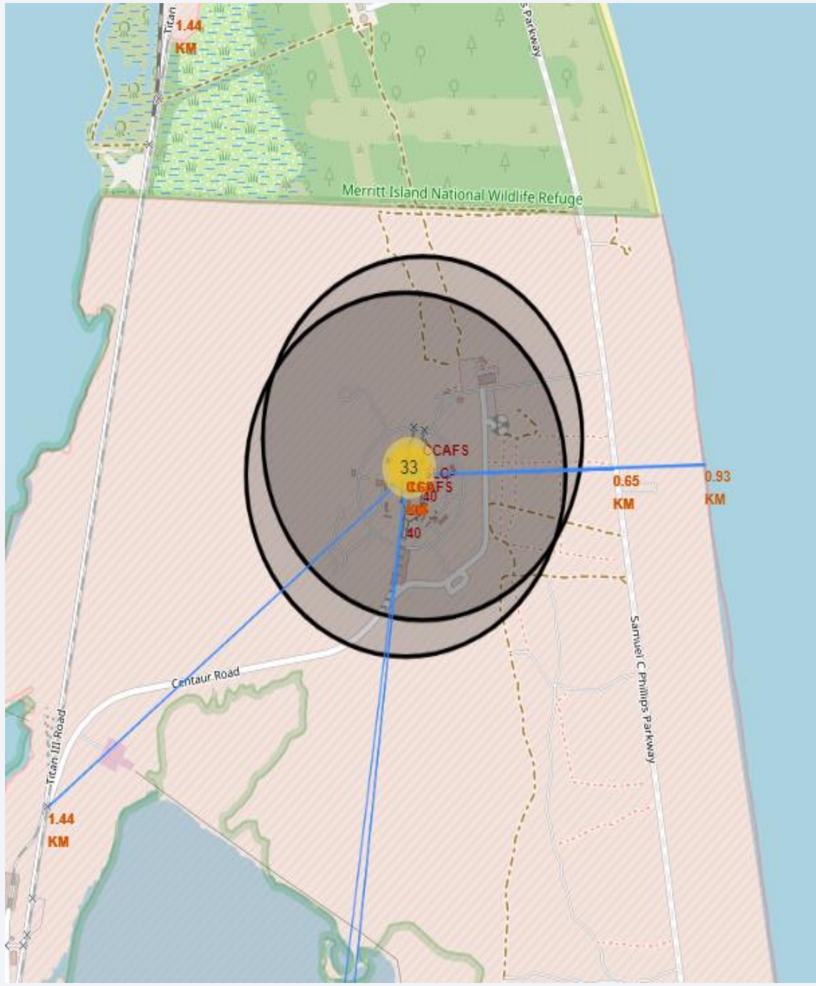
SpaceX Falcon9 – Successful and Failed Launches



- Marker clustering simplifies maps with numerous markers at identical coordinates. Using color-coded markers, it becomes easier to discern which launch sites exhibit higher success rates. Green markers indicate successful launches, while red markers denote failed launches. Notably, Launch Site KSC LC-39A exhibits a significantly high success rate.

SpaceX Falcon9 – Launch Site to proximity Distance Map



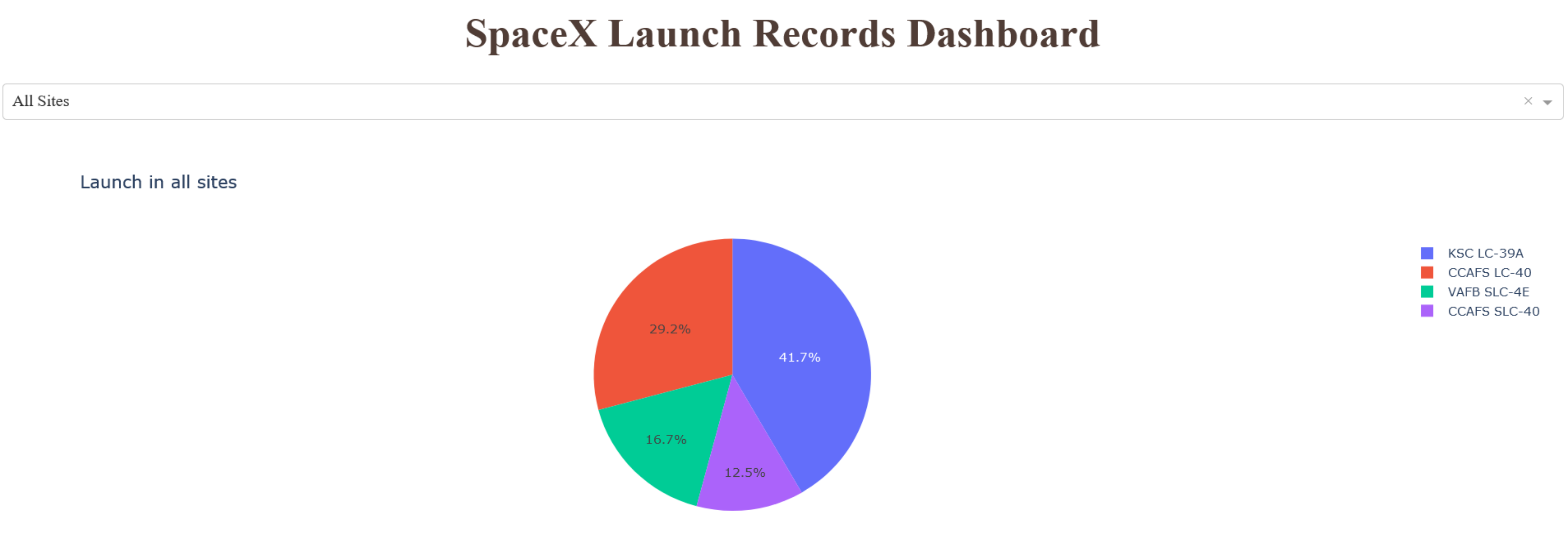


- Launch sites benefits from favorable logistical features, being close to railroads and roads, and situated at a considerable distance from populated areas. Additionally, as previously noted, its proximity to the sea facilitates rapid action and potential damage mitigation

Section 4

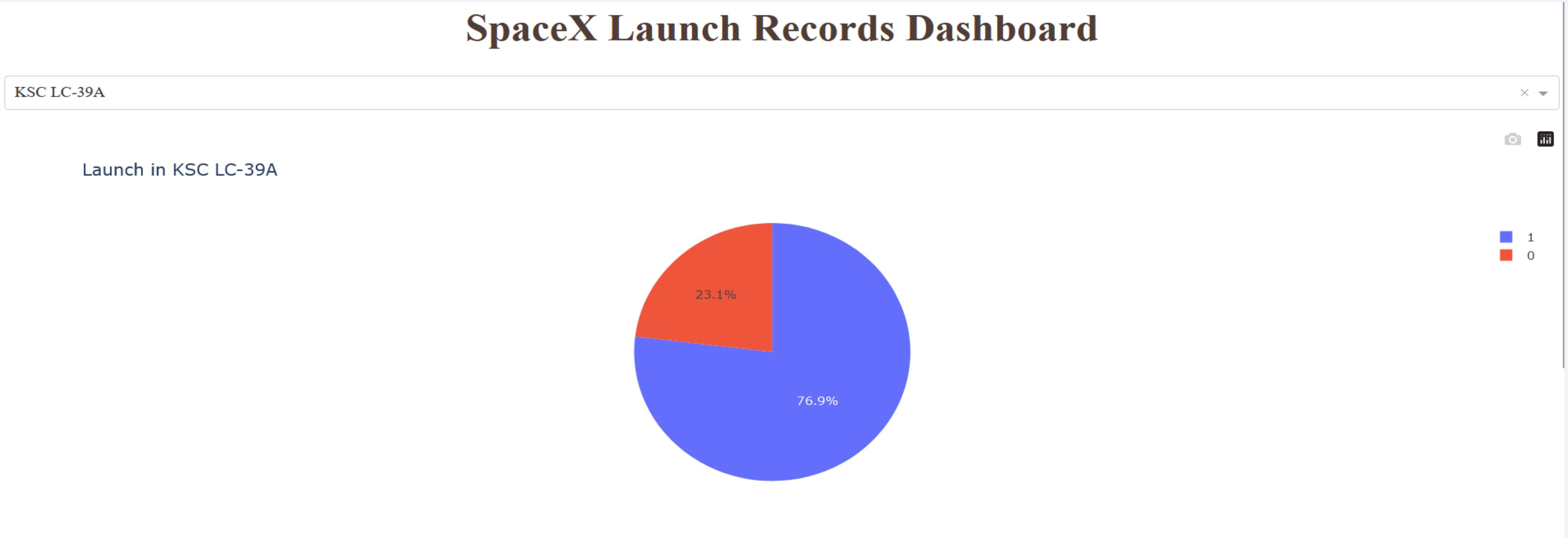
Build a Dashboard with Plotly Dash

Successful Launches by Site



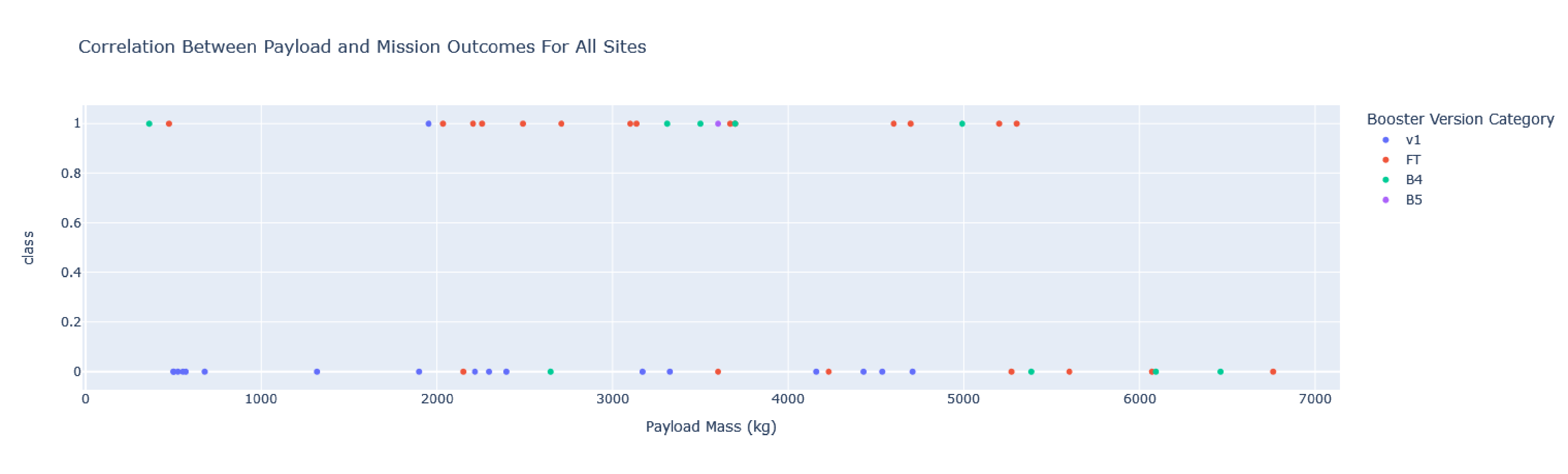
- Launch Site 'KSC LC-39A' boasts the highest rate of successful launches
- While Launch Site 'CCAFS SLC 40' records the lowest success rate for launches

Launch Success Ratio for KSC LC-39A



- KSC LC-39A launch Site has the highest launch success rate (76.9%)
- And launch failure rate 23.1%

Payload vs. Launch Outcome



- F9 Booster version with highest launch success rate is FT
- Payload range(s) with most highest launch success rate is 2000 – 5000 kg
- Payload range(s) with lowest launch success rate 0 – 2000 and 5500 – 7000

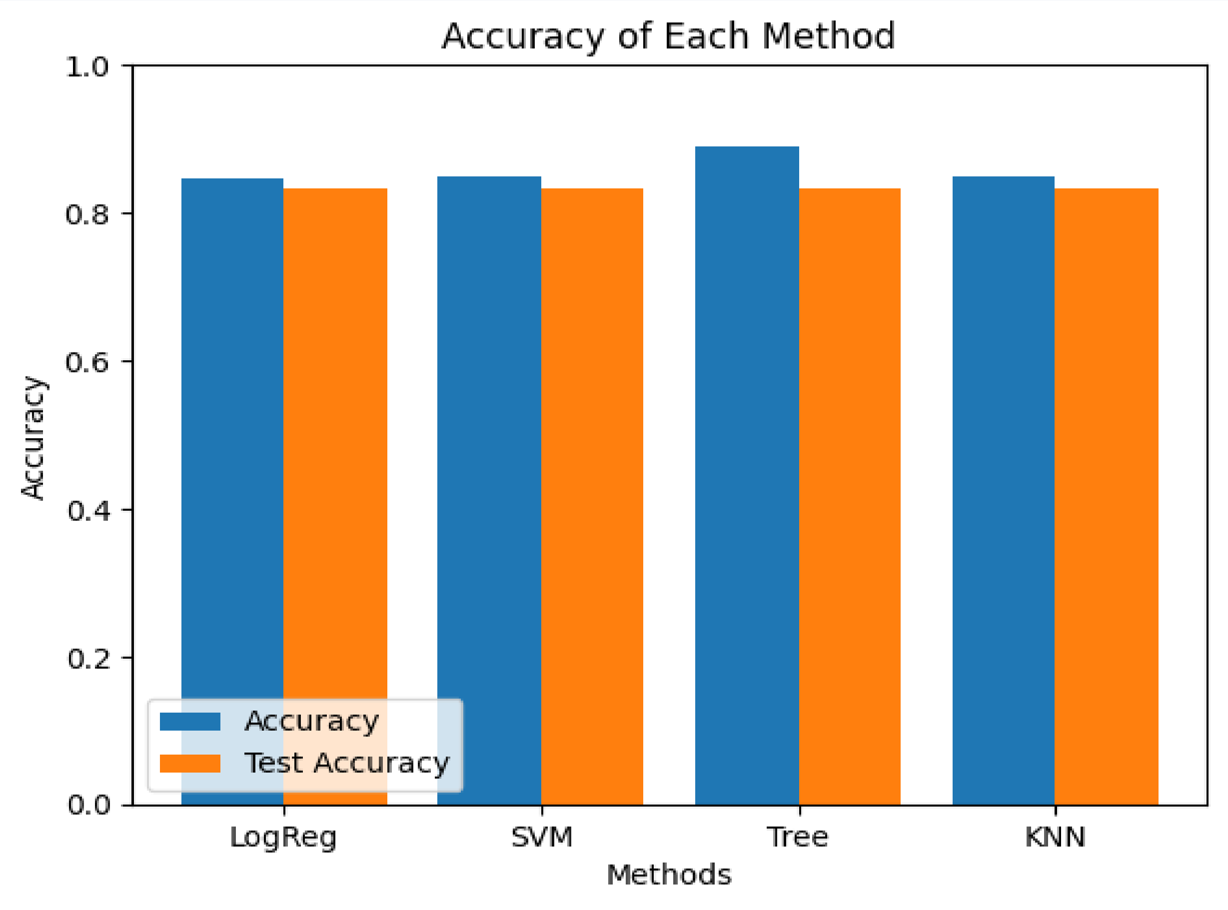
Section 5

Predictive Analysis (Classification)

Classification Accuracy

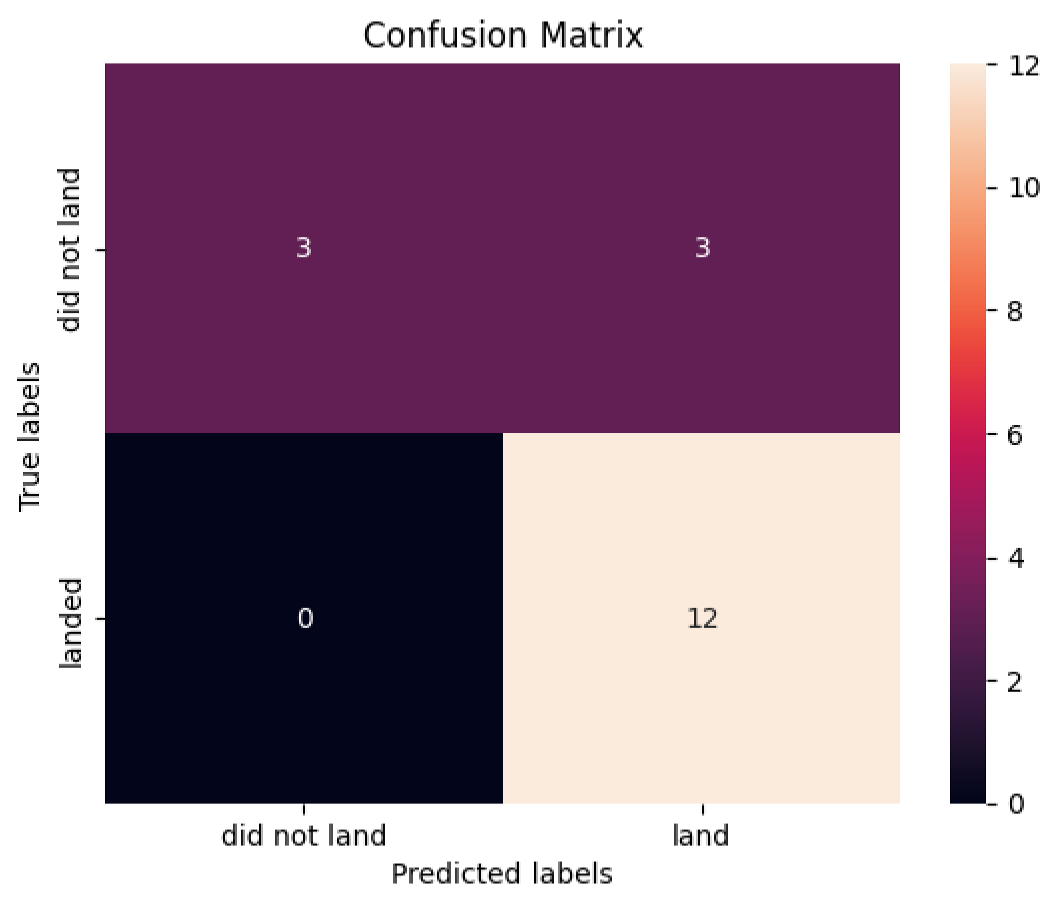
- Based on the accuracy scores and as reflected in the bar chart, the Decision Tree algorithm achieves the highest classification score, recording a value of 0.88929. The accuracy score across all classification algorithms on the test data remains consistent, each scoring 0.8333. Given the proximity of the accuracy scores for the classification algorithms and the uniformity of the test scores, expanding the dataset might be necessary to enhance model tuning with a bigger sample size.

Model	Accuracy	TestAccuracy
LogReg	0.84643	0.83333
SVM	0.84821	0.83333
Tree	0.88929	0.83333
KNN	0.84821	0.83333



Confusion Matrix

- The model is interesting due to its frequent correct predictions of the labels
- However, it incorrectly predicted mission success three times when the missions actually failed Minimizing false positives is crucial to prevent wasting millions of dollars and years of effort
- Considering a model with lower overall accuracy but higher precision is indicated
- As previously mentioned, a larger dataset is necessary for more precise evaluation and analysis



Accuracy: (TP+TN)/Total = (12+3)/18 = 0.83333
Misclassification Rate: (FP+FN)/Total = (3+0)/18 = 0.16667
True Positive Rate: TP/Actual Yes = 12/12 = 1
False Positive Rate: FP/Actual No = 3/6 = 0.5
True Negative Rate: TN/Actual No = 3/6 = 0.5
Precision: TP/Predicted Yes = 12/15 = 0.8
Prevalence: Actual yes/Total = 12/18 = 0.6667

Conclusions

- While most mission outcomes are successful, the rate of successful landings has shown improvement over time, paralleling advancements in rocket technologies; the overall launch success rate surged by approximately 80% from 2013 to 2020
- Multiple data sources were examined, leading to refined insights throughout the process
  - Nevertheless, acquiring more data would be advantageous for enhancing these insights, and undoubtedly, engineers and scientists incorporate this data into their forecasts
- Launch Site 'KSC LC-39A' achieves the highest launch success rates, whereas Launch Site 'CCAFS SLC 40' records the lowest
- Orbits such as ES-L1, GEO, HEO, and SSO report the highest success rates, with GTO orbit showing the lowest
- Launch sites are optimally positioned away from urban areas yet close to coastlines, railways, and highways
- The Decision Tree Classifier is employed to predict successful landings and boost profits, standing out as the top-performing machine learning classification model with an accuracy of roughly 89%. On test data, the accuracy reached about 83% across all models
- More comprehensive data could assist in fine-tuning the models and finding a better fit. The relationship between low and high payloads needs to be directly evaluated with the addition of more data

Thank you!